DOCUMENT RESUME

ED 262 045 TM 850 500

AUTHOR Gardner, Eric F.
TITLE How Can Tests be Misused: Let Me Count the Ways.
PUB DATE Apr 85
NOTE 11p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 1-4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Viewpoints (120)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Achievement Tests; Aptitude Tests; College Entrance Examinations; Elementary Secondary Education; Error of Measurement; Higher Education; Information Utilization; Scores; *Standardized Tests; Test Bias; *Testing Problems; *Test Interpretation; Test Manuals; Test Norms; *Test Use; Test Validity

ABSTRACT

Criticism of standardized tests is extensive, and a number of claims of test misuse have been made. Test misuse is discussed in six common circumstances. (1) Test users accept the title of a test as an accurate and complete description of the variable being measured, not having full knowledge of the test, its items, and the test manual. A cognitive test may test memory and vocabulary, but not reasoning ability. (2) The score is accepted while ignoring the error of measurement. (3) A single score is used for decision making, without considering all the factors influencing that score. (4) Test difficulty is not considered when interpreting a score, or the exact meaning of the type of score is not understood. (5) The information provided by the test score is confused with interpretations of what caused the behavior described by the score. (6) The test is biased when the predicted score is consistently too high or too low for a subgroup, or when the score does not represent an adequate measure of the ability being tested. This would occur when a low ability reader scores poorly on a science test because of the reading level of the items. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED262045

Presented at the 1985 NCME
Annual Meeting

Dr. Eric F. Gardner
Syracuse University

How Can Tests Be Misused: Let Me Count The Ways

That criticism of standardized tests is extensive and far-reaching is obvious. The reasons are less so. In many instances, especially with regard to political issues, tests provide a handy target on which to focus criticism. Tests are also criticized unjustly and sometimes appropriately for the wrong reasons. In addition, tests are used for a variety of purposes, sometimes correctly and sometimes incorrectly, that have different effects on different publics. Finally, standardized tests do have limitations and are subject to misuse. The misuse has been compounded by the decrease in training in the technical aspects of testing (construction and use) and measurement concepts and skills offered by teacher training institutions as well as by the public's lack of real knowledge about what tests can or cannot be expected to accomplish. These are some of the major reasons for the extensive criticism of tests, which must be taken into account in assessing use and misuse.

There are numerous lists of criticisms and claims of misuses of tests. These range from a list by the NEA culminating in a demand for the elimination of standardized tests to a list compiled from suggestions made by the professional staff of The Psychological Corporation. Several years ago, a graduate student at Boston College produced a list of well over sixty specific criticisms from his review of the literature and the last I knew was still adding to his list.

TM 850 500

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
Gardner, Eric F
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

When separated from such general misunderstandings, (All criticisms are not valid nor represent misuse.) most testing misuse would seem to fall into one of the following six general areas:

1. Acceptance of a Test Title for What the Test Measures

There is a tendency for unsophisticated test users to accept the name assigned to a test as an accurate and complete description of the variable being measured. Since titles must be brief, they cannot convey all that the user needs to know about the kind of behavior to be measured. All tests are open to this kind of uncritical abuse. Since there are so many facets of cognitive ability, it is obvious that no test can be an adequate measure of them all. Only full knowledge of the test and its items can reveal what is being measured. Furthermore, the testing situation may completely change the expected behavior on the test. If a non-English speaking or blind pupil is given an "aptitude" test in printed English, it obviously doesn't measure any aspect of "aptitude" or "intelligence" except lack of knowledge of English or lack of vision. A test producer's claims for an achievement test or an aptitude test do not mean that it will function as such in all circumstances with all pupils. Failure to examine the manual and the items carefully in order to know the specific aspects of cognitive ability to be tested (memory, vocabulary, type of reasoning, etc.) can result in misuse by virtue of selecting an inappropriate test for a particular purpose or situation.

2. Ignoring the Error of Measurement in Test Scores

Every test score contains an error of measurement. It is a misuse of any test score or any observation to accept it as a fixed, unchanging index containing no error. It is inappropriate to assert or lead a user to believe that an individual's observed score gives his "true" performance on the general domain about which inferences are to be made. The best that can be done is to estimate experimentally the standard error of measurement and then use that value to set up a confidence interval within which one can state that the "true" score falls within that band with a specific probability of being wrong. That 1) we cannot accept an achievement score of 45 as a precise measure, 2) we must accept a range of scores, and 3) we must then expect to be wrong a certain proportion of the time in asserting that the true score falls within a specific confidence interval does not mean that the achievement test does not furnish useful data. It does mean that the test score is being misused if knowledge of the size of the errors of measurement is not used in interpreting the score. In the case of most standardized test scores, the magnitude of the errors is made explicit, not hidden or unknown.

3. Use of a Single Test Score for Decision Making

Misuse of tests occurs when scores are not considered and interpreted in the full context of the various elements that characterize pupils, teachers, and the general educational environment involved, for a test score represents only a sample from a limited domain and

does not include the variety of factors that might influence that score. For example, in decisions determining admission to college, SAT scores should not be used in isolation and are in fact usually considered along with the pupil's high school record and other relevant data such as teacher's or supervisor's recommendations concerning motivation, leadership ability, creativity, involvement in extracurricular activities, etc. All of these characteristics can then be evaluated against the student's socioeconomic background, along with consideration of any social obstacles or unusual physical demands required of the student to reach his current educational level (Fruen 1978).

4. Lack of Understanding of Test Score Reporting

There is substantial misunderstanding, not just among laymen, but also among many educators, of the meaning of test scores. Most people believe that they understand the meaning of a raw score or of that particular raw score converted to a percent of items answered correctly, as in the case of many criterion-referenced tests. However, even in this most elementary illustration, more is involved than a single number indicates. Forty-five items answered correctly out of fifty easy items has a substantially different meaning than forty-five items answered correctly out of a sample of fifty very difficult items from the same domain.

The interpretation of a raw score converted to a percentile score causes even more problems. The statement that "In a norm-referenced test half the pupils must fail" is a good illustration. A percentile

score of 20, by itself, represents neither a good nor a poor performance. It merely indicates that among the group used as a frame of reference this score was higher than that reached by 20 percent of its members. If the group were of high ability or had unusual skills, a percentile rank of 20 might indicate an excellent or even remarkable performance.

The misinterpretation of grade equivalents is even more common. A grade equivalent is the score that was exceeded by 50 percent of the normative group and not attained by the other 50 percent of the group at the specific time when the test was given. It does not represent a standard to be attained. It does not represent the grade in which the pupil should be placed.

To compensate for the decreasing emphasis on test construction and test interpretation in teacher training institutions, there have been efforts by the National Council on Measurement in Education and other organizations to provide workshops and reading material on measurement issues. Both parents and professional educators stand to benefit since both are involved in the misuse of testing based on misinterpretation of scores.

5. Attributing Cause of Behavior Measured to the Test

It is common, especially for critics of testing, to confuse the information provided by a test score with interpretations of what caused the behavior described by the score. A test score is a numerical description of a sample of performance at a given point in time. Why an individual obtained a particular score is an entirely separate issue and the score itself does not give an answer to this question.

Claiming that it does, whether intended as a positive attribute or a criticism, is tantamount to test misuse. Furthermore, no statistical manipulation of test data, even though combined with the best additional data, will permit more than probabilistic inferences about causation or future performance. The current reports (Beaton et al. 1977; College Entrance Examination Board 1977) on the decline of SAT scores is an excellent example of the difficulty in ascribing causation to known performance. The charge given the researchers by the investigating panel was to explain the causes of the drop in SAT scores. They were able to describe the drop and offer changes in test populations as a plausible partial explanation for the initial drop but could only speculate on the effect of other variables and the reasons for the continued drop.

6. Bias

There is an enormous literature on bias associated with tests, which bears chiefly on two aspects of test use or misuse: bias in a selection process using test scores; and test bias or the bias ascribed to a particular test or test item (Gardner 1978). The first situation is one in which the test is used to predict a criterion. A test is biased for members of a subgroup of the population if, in the prediction for which the test was designed, the predicted score is consistently too high or too low for members of that subgroup. The second involves bias in the test itself, in the sense that the score does not represent an adequate measure of the specific ability being tested.

A number of approaches have been proposed as solutions to the problem of bias in selection, ranging from a least squares regression

method to various types of procedures based on specific ratios and probabilities. Part of the difficulty with statistical selection models such as the Constant Ratio or Conditional Probability Model (or, for that matter, the Regression of Equal Risk Models) is that users often expect to obtain a purely technical resolution to an issue that involves value judgments. Values are implicit in many models, but they need to be dealt with explicitly. Decision theory provides a framework for doing this, and that is the approach advocated by Peterson and Novick (1976). However, although it provides a framework for making a decision, it doesn't automatically provide the decision. The answer requires a judgment, which involves the background and expertise of the decision maker and hence implies potential bias. It is important that the point where judgments or values enter the process or are incorporated in the statistical process be made explicit and understood.

The literature on test bias or set- and culture-free tests is far more voluminous than that on bias in selection. The use of the word bias in this context has even more connotations than it does when associated with selection, as the possibility of bias is present in all tests. In some instances bias does exist; it may fail to be claimed, or it may be claimed for the wrong reasons. But more often when it doesn't exist at all, it may be claimed for irrelevant reasons. For example, as Ebel has pointed out (1976) some standardized educational achievement tests have been attacked for their alleged bias against cultural minorities and pupils of low socioeconomic status because such pupils tend to score lower.

Lower scores alone do not signify bias in an achievement test. If they did, then every spelling test would be biased against poor spellers, every vocabulary test against persons who had poor vocabularies, and every shorthand test against persons who had never learned shorthand. An achievement test is designed to measure the competence of pupils in specifically defined areas under conditions described in the manual. The score of each pupil indicates how adequately the questions defining the domain tested were answered. For a test to have satisfactory content validity, it is important that aspects of the variable the test was designed to measure not be deleted in the name of bias.

In contrast, a test designed to measure achievement in elementary science, in which the items are presented in language and syntax far above the level of the pupils tested, is biased especially against the poor readers. A distinction must be made here between that aspect of reading difficulty based on science vocabulary that was part of the curriculum and other vocabulary that was not. The score on the test should measure knowledge of science, including appropriate scientific vocabulary, but not general reading ability. This is a case of bias introduced by complex reading material, which can be recognized and corrected.

Recently, the lowest age grade level of a diagnostic reading test was rejected by a state committee because of sex bias. The primary criticism was that the items portrayed males and females in mostly traditional roles. The criticism was not made of the higher level tests. Here is an illustration where trade-offs are necessary.

If the objective of the test was to promote the concept of sexual equality and to teach that sex is irrelevant to a person's roles in life, then it is clear that the test was counterproductive for that purpose. If the purpose was to present stimuli familiar to very young children just beginning to learn to read, then females in the roles of mother and teacher and males as father, mail carrier and fire fighter are reasonable, as the aim should be to avoid distraction from the reading task.

Conclusion

Concern about the misuse of tests is justified. The heat of the controversy testifies to the interest of the public in one of this country's most widely used educational tools. A missing element, however, in almost all the debates on the use and misuse of tests is a reasonable degree of humility. Such statements as "I don't know," or "We need to find out about that" are rare. Without some recognition of the fallibility of humans and human minds, profitable, meaningful, and intelligent dialogue is impossible.

References

Fruen, M. The use of tests in admissions to higher education, Measurement in Education, 9(Fall): 4, 1978.

Beaton, A.E., Hilton, T.L., and Schrader, W.B. Changes in the Verbal Abilities of High School Seniors, College Entrants, and SAT Candidates between 1960 and 1972. Princeton, New Jersey: Educational Testing Service, 1977.

College Entrance Examination Board. On Further Examination. A Report of the Advisory Panel on the Scholastic Aptitude Test Score Decline. New York: College Entrance Examination Board, 1977.

Gardner, E.F. Bias, Measurement in Education, 9(Summer): 3, 1978.

Peterson, N.S. and Novick, M.R. An evaluation of some models for culture fair selection, Journal of Educational Measurement, 13:3-29, 1976.

Ebel, R.L. A paradox of educational testing, Measurement in Education, 7:4, 1976.